Supplementary Material For ICCV 2021 Submission #5850

1. Overview

In this supplementary file, we will give more detailed information as mentioned in the main paper. We add the word "Extended" in the section title to mean that the section contains extended information about the corresponded section in the main paper.

2. The H2O Handover Dataset, Extended

2.1. The Video Recording Setup

Alongside this supplementary file, we attach a video demo to show how we record the handover process in action. The main view is recorded by a Redmi K30 pro smart phone, which is not one of the 5 RGB-D cameras. We only use Redmi to show how a director works during recording.

As for the RGB-D stream, we select 3 of all 5 views to show up in the video, which is only for demonstration. When building the H2O dataset, all 5 RGB-D streams are recorded and stored in the dataset.

2.2. Task-Oriented Intention

We pre-define common tasks for all 30 objects, and the volunteers should pick one for task-oriented handover. The tasks are borrowed some from ContactPose [?], but we also add a few if other high-priority options exist in the daily life. The full list can be referred to Table ??. Note, the "casual" is default for all the objects, thus we don't mention it in the Table ??. The object name follows the convention of YCB object [?] and ContactPose [?].

Object Name	Task List
ContactPose	
apple	eat; wash
banana	peel
binoculars	see through
bowl	drink from; catch sth.; wash
camera	take a picture
cell phone	talk on
cup	drink from; wash
eyeglasses	wear
flashlight	turn on; sweep
hammer	hit a nail; push; sweep
headphones	wear
knife	cut
light bulb	screw in a socket
mouse	use to point and click
mug	drink from; wash
pan	cook in
PS controller	play a game with
scissors	cut with
stapler	staple
toothbrush	brush teeth
toothpaste	squeeze out toothpaste
Utah teapot	pour tea from; wash
water bottle	open
wine glass	drink wine from
YCB	
Scrub Cleanser bottle	squeeze; screw; pour
French's Mustard bottle	squeeze; screw; pour
Spam Potted Meat can	open the lid
Power Drill	pull the trigger; hold
Scissors	cut with
Large marker	write; pull off the hat

Table 1. The object and its intended tasks when requested.

2.3. Image Split and Statistics

We have mentioned the essential part of the dataset statistics in the main paper. In this section, we will add some details about the dataset.

Handover-wise Statistics In H2O dataset, we have recorded 6K handover process which contains 40 giver-receiver pairs passing over 30 objects under 5 cameras. For each handover video, we split them into 3 stages (namely pre-handover, physical handover and post-handover) and result in a total of 18K video clips.

Stage-wise Statistics In H2O dataset, among all 5M image frames, the pre-handover stage contains 2.5M frames, the physical handover stage contains 1M frames and the post-handover stage 1.5M frames. The pre-handover stage can be used for tasks such as hand pose estimation, hand-object pose estimation (interaction reconstruction), trajectory prediction etc. The physical handover stage can be used for analyze how two hands interact with one object, such as the proposed task Receiver Grasp Prediction. The post-handover stage contains rich task-oriented grasp information. Different stage of handover process has different features in itself, thus can support multiple tasks.

When spliting the dataset for hand-object interaction and receiver grasp prediction, we randomly select 500 video clips in the pre-handover subset for validation and other 500 for testing. We don't seek the method can be generalized to unseen object, but unseen hand configuration. Thus all the objects can be seen during training and testing.

3. RGPNet, Extended

3.1. Hand-Object Area Extraction

We first detect the hand-object region with a pre-trained detector [?], which predicts both the hand and the object bounding box. Then we merge the hand-object bounding box pairs to form a larger bounding box if the two bounding boxes overlap larger than 10%, which is set empirically. The merged bounding box is then enlarged 1.2 times on both width and height. To note, if the object or the hand is not detected, due to occlusion, we just enlarge the bounding box of the hand with the same ratio.

4. Experiments & Results, Extended

4.1. Results on H2O-Syn Dataset

When training on H2O-Syn dataset, we change the background with COCO [?] dataset. The background is randomly selected. We report both results on the test split of H2O and H2O-Syn dataset.

Hand-object Interaction Reconstruction For hand-object interaction reconstruction task, we adopt the same baseline methods as in the main paper. The results are reported in Table ??. When train and test in the same domain (H2O-Syn), the object error and hand error is at the same level of the H2O results which are reported in the main paper. It is as expected, since the object pose and hand pose is the same. On the other hand, it is noteworthy that the object pose is affected more than the hand pose when suffering the domain gap between H2O-Syn and H2O.

Receiver Grasp Prediction For receiver grasp prediction task, we also implement both RGPNet and RGPNet with grasp type prediction. The results are reported in Table ??. Since the receiver grasp prediction depends on the result of hand-object interaction reconstruction, due to the domain gap between the H2O-Syn and H2O dataset, especially the prediction of object pose, the grasp score and interpenetration error suffer a rather large loss.

Test Split	Method	Object error	Hand error
H2O-Syn	Tekin et al. [?]	27.6	18.7
	Hasson et al. [?]	26.3	19.5
H2O	Tekin et al. [?]	31.3	19.8
	Hasson et al. [?]	29.4	22.6

Table 2. Hand-object reconstruction results. Training on H2O-Syn pre-handover subset, testing on H2O-Syn and H2O pre-handover subset respectively. The errors are measure in mm.

Test Split	Method	Grasp Score ↑	Interp. ↓
H2O-Syn	RGPNet	0.63	22
	RGPNet + grasp type	0.66	21
H2O	RGPNet	0.46	33
	RGPNet + grasp type	0.52	31

Table 3. Receiver grasp prediction results. ↑ means the the higher the better; ↓ means the lower the better.

4.2. Implementation Details

In this section, we will describe the detailed architecture of the RGPNet.

For the Generator, the dimension of the global feature $\mathcal{F}$ is 256. We have also tried with other numbers such as 128, 512, but found no significant difference in practice. The following 4 branches of prediction network (denoted as "FC") shares the same architecture which has 3 fully connected layers. The input and output dimension for each layer is (256, 512), (512, 1024), (1024, K) respectively, where K is different for different output. For example, $K = 4$ for predicting ΔR_w, $K = 3$ for predicting Δt_w, $K = 10$ for predicting $\Delta\beta$, and $K = 45$ for predicting $\Delta\theta$.

For the Discriminator, it is a 3-layer fully connected network. The input and output dimension for each layer is (124, 512), (512, 1024), (1024, 2). The input dimension 124 is resulted from sum of the ΔH ($62 = 4 + 3 + 10 + 45$) multiplying 2.

4.3. Variant RGPNet with Grasp Type Prediction

When grasp type [?] is predicted, the RGP-GAN need a little adaptation for grasp type prediction. Specifically, alongside the global feature $\mathcal{F}$ predicted from the ResNet-18 [?], we also make a prediction of the grasp type, which is one-hot encoded and optimized by a cross-entropy loss. Then, the predicted hand is reformulated as $H = H'_{pre} + H_o + \Delta H$, where $H_o = \{\beta_o, \theta_o, 0\}$ is the default hand configuration for each grasp type, and $H'_{pre} = \{0, 0, P_{w,pre}\}$. In other words, when predicting with grasp type, the receiver's hand configuration is regarded as a relative transformation from the default grasp template, and the hand wrist pose is a relative transformation from the giver's hand wrist.

References

[1] Samarth Brahmbhatt, Chengcheng Tang, Christopher D Twigg, Charles C Kemp, and James Hays. Contactpose: A dataset of grasps with object contact and hand pose. arXiv preprint arXiv:2007.09545, 2020. 1

[2] B. Calli, A. Singh, A. Walsman, S. Srinivasa, P. Abbeel, and A. M. Dollar. The ycb object and model set: Towards common benchmarks for manipulation research. In 2015 International Conference on Advanced Robotics (ICAR), pages 510–517, 2015. 1

[3] T. Feix, J. Romero, H. B. Schmiedmayer, A. M. Dollar, and D. Kragic. The grasp taxonomy of human grasp types. IEEE Transactions on Human-Machine Systems, 46(1):66–77, 2016. 2

[4] Yana Hasson, Bugra Tekin, Federica Bogo, Ivan Laptev, Marc Pollefeys, and Cordelia Schmid. Leveraging photometric consistency over time for sparsely supervised hand-object reconstruction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 571–580, 2020. 2

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770–778, 2016. 2

[6] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In European conference on computer vision, pages 740–755. Springer, 2014. 2

[7] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. arXiv preprint arXiv:1506.01497, 2015. 2

[8] Bugra Tekin, Federica Bogo, and Marc Pollefeys. H+ o: Unified egocentric recognition of 3d hand-object poses and interactions. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 4511–4520, 2019. 2